The Rich Transcription 2006 Evaluation Overview and Speech-To-Text Results

http://www.nist.gov/speech/tests/rt/rt2006/spring/

Jonathan Fiscus, John Garofolo, Jerome Ajot, Martial Michel

May 3, 2006

Rich Transcription 2006
Spring Meeting Recognition Workshop
at MLMI 2006

Overview

- Rich Transcription Evaluation Series
- RT-06S Evaluation
  - Audio input conditions
  - Corpora
  - STT Evaluation task and results
- Conclusion/Future

The Rich Transcription Task

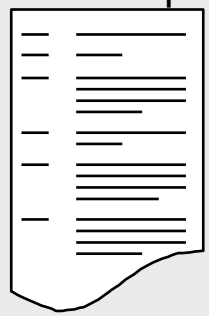
Human-to-Human Speech

Component Recognition Technologies

RICH TRANSCRIPTION
Speech-To-Text + METADATA

Multiple Applications

Readable Transcripts



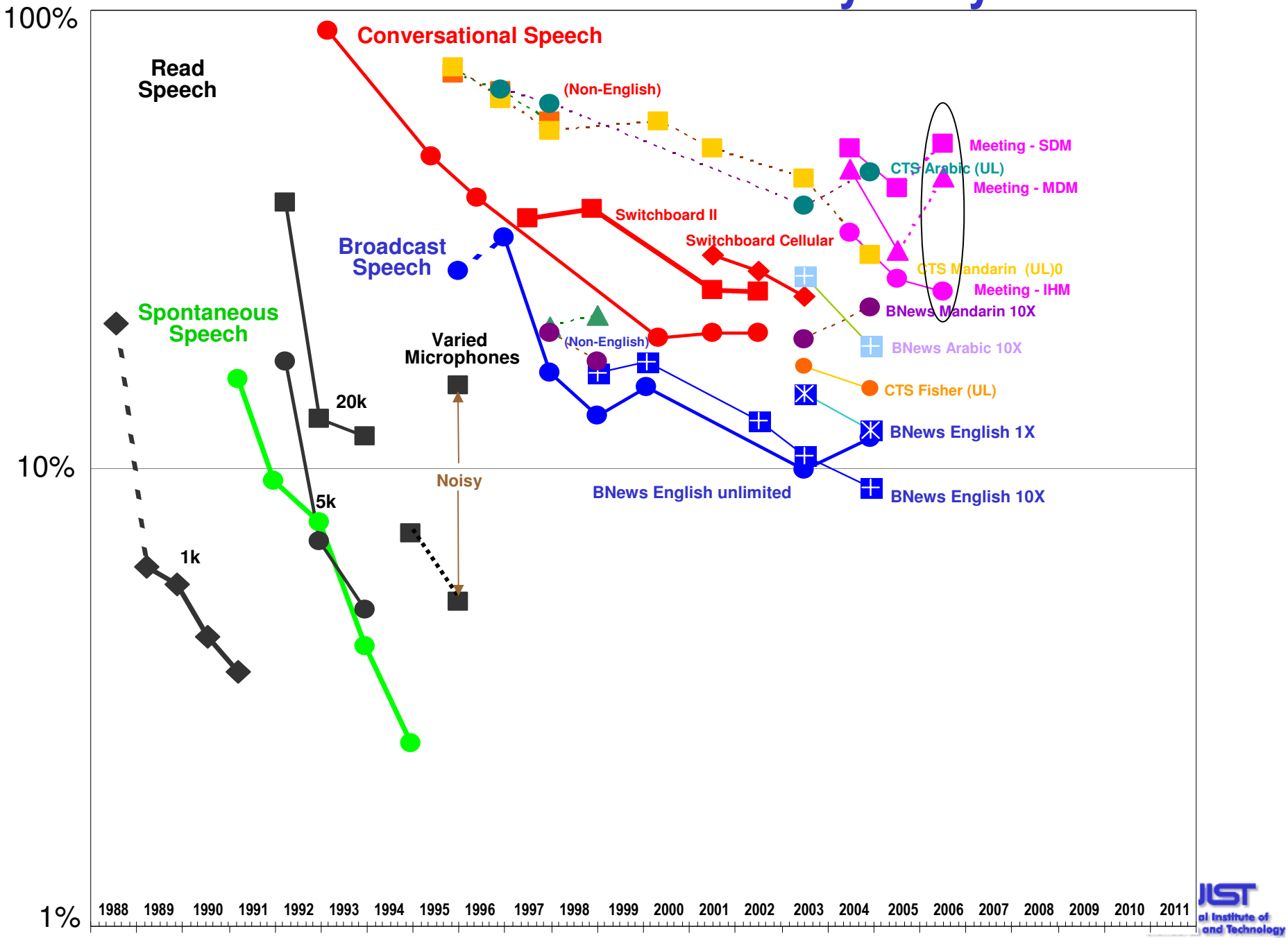
Smart Meeting Rooms
Translation
Extraction
Retrieval
Summarization
...

NIST National Institute of Standards and Technology

Rich Transcription Evaluation Series

- Goal:
  - Develop recognition technologies that produce language content representations (transcripts) which are understandable by humans and useful for downstream processes.
- Domains:
  - Meeting Room speech
  - Broadcast News (BN)
  - Conversational Telephone Speech (CTS)
- Parameterized “Black Box” evaluations
  - Evaluations control input conditions to investigate weaknesses/strengths
  - Sub-test scoring provides finer-grained diagnostics

NIST STT Benchmark Test History – May. '06



Collaborations

- Augmented Multiparty Interaction (AMI) Program
- Computers in the Human Interaction Loop (CHIL) Program
- Classification of Events, Activities, and Relationships (CLEAR) Workshop
  - NEW EVALUATION

CLEAR Evaluation

- Classification of Events, Activities, and Relationships Evaluation Program and Workshop series
  - Focus is on multi-modal technologies for human activity and interaction analysis
  - Collaboration across programs (like RT)
    - CHIL, VACE in 2005
    - also AMIDA and PETS in 2006
  - Combining video and audio processing and other modalities
  - 23 evaluation tasks supported in CLEAR ‘06
    - Utilized same datasets as RT for some tasks

2006 CLEAR Evaluations

		Source Data					
Task	Sub-condition	Multi-Site Conference Meetings	Seminar Meetings	Surveillance	Broadcast News	Studio Poses	UAV
3D Sing Per Track	Video		CHIL				
	Audio		CHIL				
	Audio+Video		CHIL				
3D Multi Per Track	Video	CHIL					
	Audio	CHIL					
	Audio+Video	CHIL					VACE
2D Multi Per Track	Video			VACE			
2D Face Det	Video	VACE, CHIL	CHIL		VACE		
2D Face Track	Video	VACE			VACE		
Person ID	Video		CHIL				
	Audio		CHIL				
	Audio + Video		CHIL				
Head Pose Est	Video		CHIL			CHIL	
Vehicle Track	Video			VACE			VACE
Acoustic Event Detection	Audio		CHIL				
Environment Class	Audio		CHIL				

NIST National Institute of Standards and Technology

RT-06S Evaluation Tasks

- Focus on core speech technologies – extracting speech content from audio modality
  - Speech-To-Text Transcription
    - Transcribe the spoken words
  - Diarization “Who Spoke When”
    - Identify the number of participants in each meeting and create a list of speech time intervals for each participant
  - Diarization “Speech Activity Detection”
    - Identify the time intervals where one or more people are talking

Six System Input Conditions

- Distant microphone conditions
  - Multiple Distant Microphones (MDM)
    - Three or more centrally located table mics
  - Multiple Source Localization Arrays (MSLA)
    - Inverted “T” topology, 4-channel digital microphone array
  - Multiple Mark III digital microphone Arrays (MM3A)
    - Linear topology, 64-channel digital microphone array
  - All Distant Microphones (ADM)
- Contrastive microphone conditions
  - Single Distant Microphone (SDM)
    - Center-most MDM microphone
    - Gauge performance benefit using multiple table mics
  - Individual Head Microphones (IHM)
    - Performance on clean speech
    - Similar to Conversational Telephone Speech
      - One speaker per channel, conversational speech

Training/Development Corpora

- ICSI Meeting Corpus
- ISL Meeting Corpus
- NIST Meeting Pilot Corpus
- Topic Detection and Tracking Phase 4 (TDT4) corpus
- Fisher English conversational telephone speech corpus
- CHIL ’05 development test set
- CHIL ’06 development test set
- AMI development data
- Rich Transcription 2004 Spring (RT-04S) Development & Evaluation Data
- Rich Transcription 2005 Spring (RT-05S) Evaluation Data

RT-06S Evaluation Corpora

- Two meeting sub-domains
  - Conference Room
    - Multi-site cross-program data collection effort
  - Lecture Room
    - Multi-site CHIL program data collection effort
    - Lecture data further divided into two categories:
      - Seminars
      - Interactive Seminars
- Evaluation corpora used by CLEAR

RT-06S Evaluation Test Corpora: Conference Room Test Set

- Goal-oriented small conference room meetings
  - Group meetings and decision-making exercises
  - Meetings involved 4-9 participants
- 162 minutes – Ten excerpts, each eighteen minutes in duration
  - Six sites donated two meetings each:
    - Carnegie Mellon Univ., Edinburgh Univ., IDIAP (donated, but not used), NIST, TNO, and Virginia Tech (VT)
  - Similar test set construction used for RT-05S evaluation
  - Transcribed by the LDC
- Microphones:
  - All participants wore head microphones
  - Microphones were placed on the table among participants
  - AMI meetings (Edinburgh, IDIAP, and TNO) included an 8-channel circular microphone array on the table

RT-06S Evaluation Test Corpora: Lecture Room Test Set

- Technical lectures in small meeting rooms
  - Educational events where a single lecturer is briefing an audience on a particular topic
- 190 minutes – 38 excerpts from 26 lectures
  - Two styles of lectures:
    - Seminar Lectures: One lecturer, large audience (between 4 and 15 people) (120 minutes)
    - Interactive Seminars: One lecturer, small audience (usually 4, sometimes more people) – 70 minutes
- Data collected at
  - AIT, IBM, ITC, Karlsruhe University, UPC
- Sensors:
  - Seminar Lectures:
    - Lecturer wore head mic, variable number of audience members wore head mics
  - Interactive Lectures:
    - Lecturer wore head mic, all audience members wore head mics
  - Microphones were placed on the table among participants
  - Inverted 'T' source localization array mounted on walls
  - Mark III mounted on the wall opposite the lecturer

Scoring/Data Problems

- Performance for 4-person speech is actually worse than reported
  - a newly discovered reference problem caused some 3-person speech segments to be scored as 4-person speech segments

- Conference Room Data
  - TNO's distant microphone data was found to be corrupt and therefore removed from distant mic scoring
- Lecture Room Data
  - American vs. English spellings (both systems and references)
  - Some IHM Channel-to-Speaker ID correspondences are incorrect and need to be re-checked
  - Three segments were transcribed for the wrong time and need to be added back into the scoring

RT-06S Evaluation Participants

Site ID	Site Name	Evaluation Task		
		STT	SPKR	SAD
AIT	Athens Information Technology		X	X
AMI	Augmented Multiparty Interaction Program	X	X	X
IBM	IBM	X		X
ICSI/SRI	International Computer Science Institute and SRI International	X	X	X
INRIA	Institut National de Recherche en Informatique et en Automatic			X
ITC-irst	Center for Scientific and Technological Research			X
KU	Karlsruhe University (UKA)	X		
LIA	Laboratoire Informatique d'Avignon		X	X
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur	X	X	X
UPC	Universitat Politècnica de Catalunya			X

Speech-To-Text (STT) Task

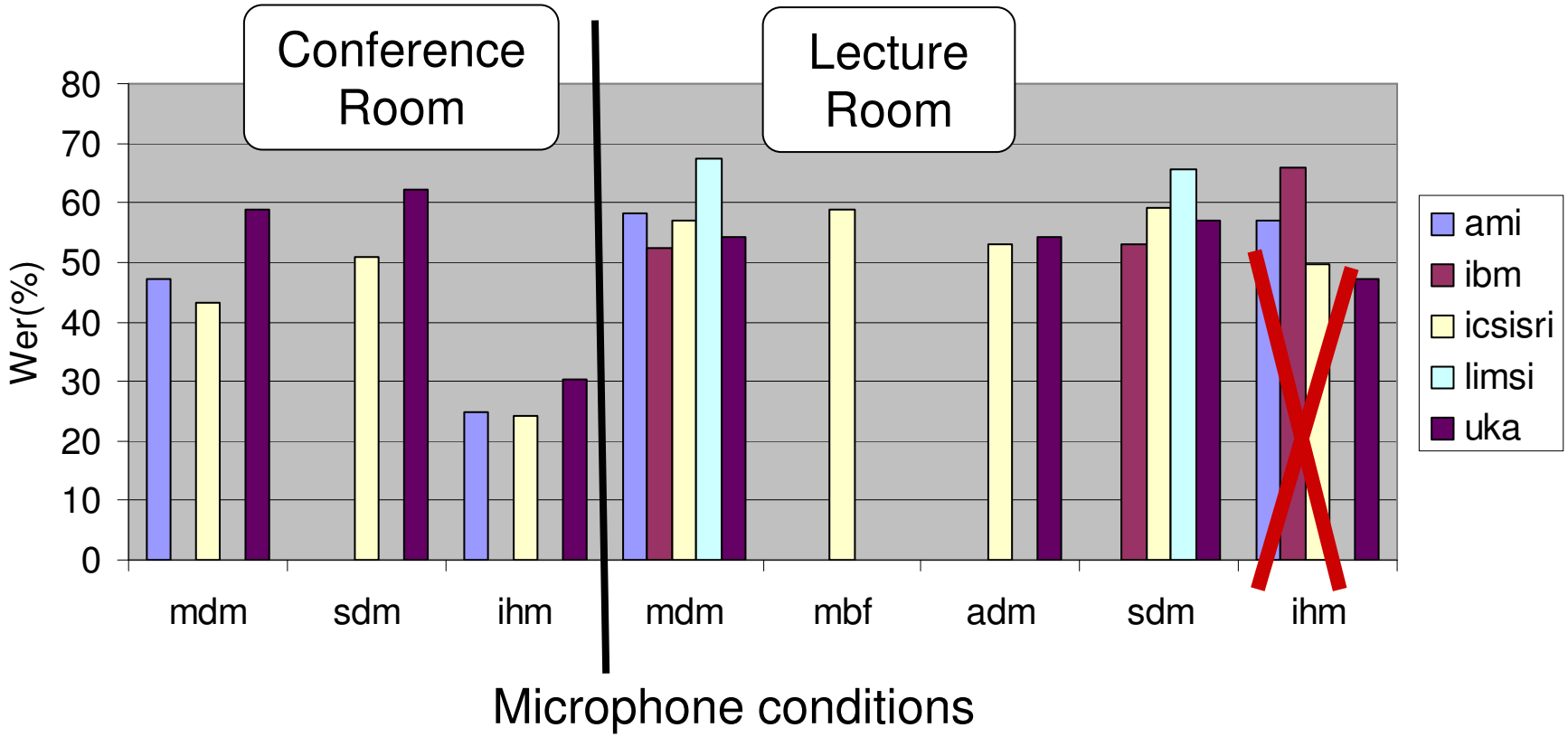
- Task definition
  - Systems output a single stream of time-tagged word tokens
- Several input conditions:
  - Conference Room: **MDM(primary)**, SDM, ADM, IHM
  - Lecture Room: **MDM(primary)**, MM3A, MBF, ADM, SDM, IHM
- Participating sites:
  - Conference Room: AMI, ICSR/SRI, UKA
  - Lecture Room: AMI, IBM, ICSI/SRI, LIMSI, UKA

STT System Evaluation Method

- Primary metric
  - **Word Error Rate (WER) -** ratio of inserted, deleted, and substituted words to the total number of words in the reference
    - System and reference words are normalized to a common form
    - System words are mapped to reference words using a word-mediated dynamic programming string alignment program
- Systems were scored using the NIST Scoring Toolkit (SCTK) version 2.1.3
  - Handles simultaneous speech
    - Periods with up to 4 overlapping speakers evaluated for Distant microphone conditions
  - Two supported STT evaluation paradigms
    - Single Stream STT system output to Multi Stream References
    - Multi Stream STT system output to Multi Stream References
    - LREC 2006: Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech

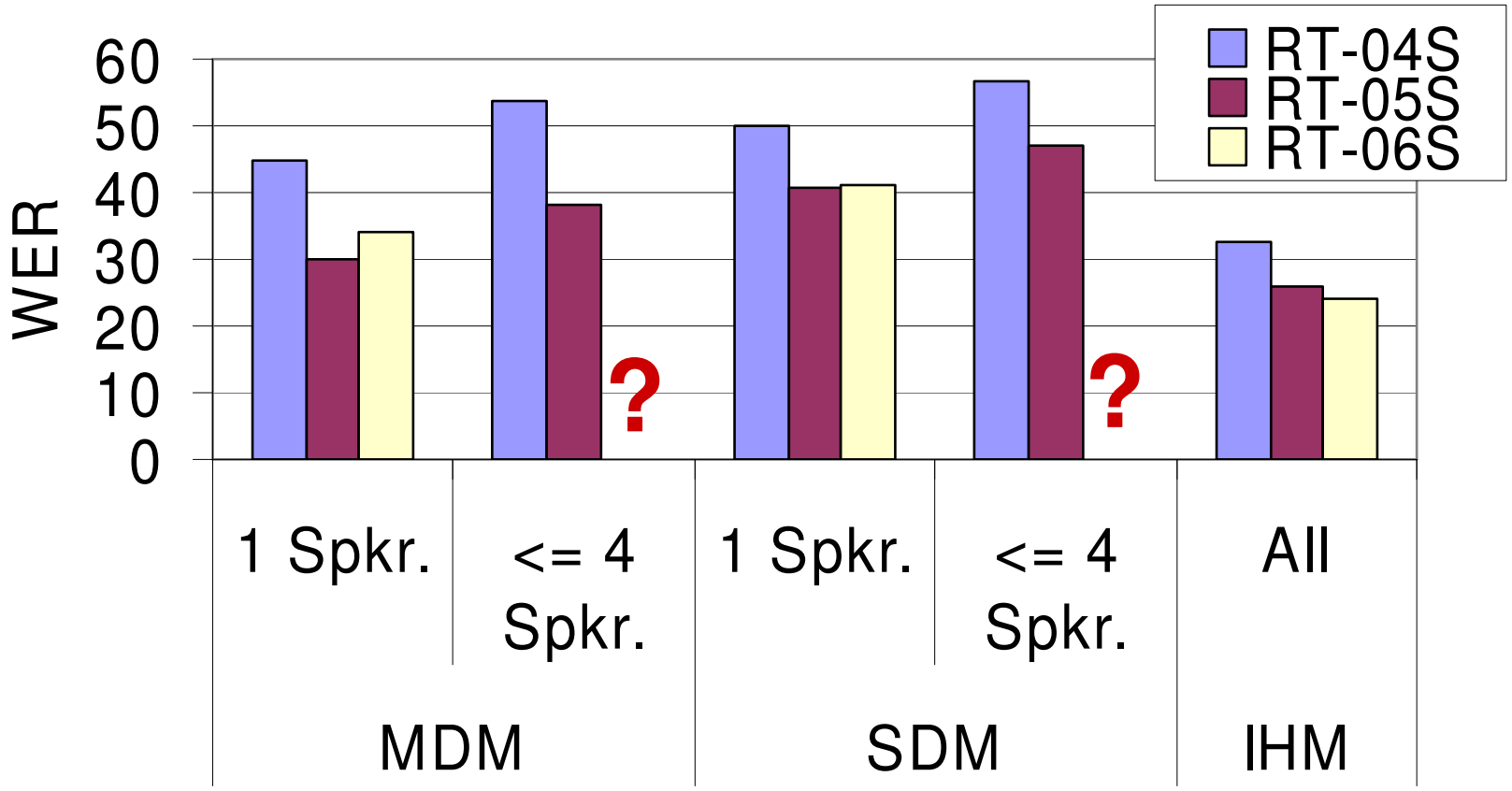
NIST
National Institute of Standards and Technology

RT-06S STT Primary System Results (Current Overlap<=4 Results)



- Conference MDM error rates are higher than last year
- Lecture IHM results will be revised

Historical STT Performance in the Conference Meeting Domain

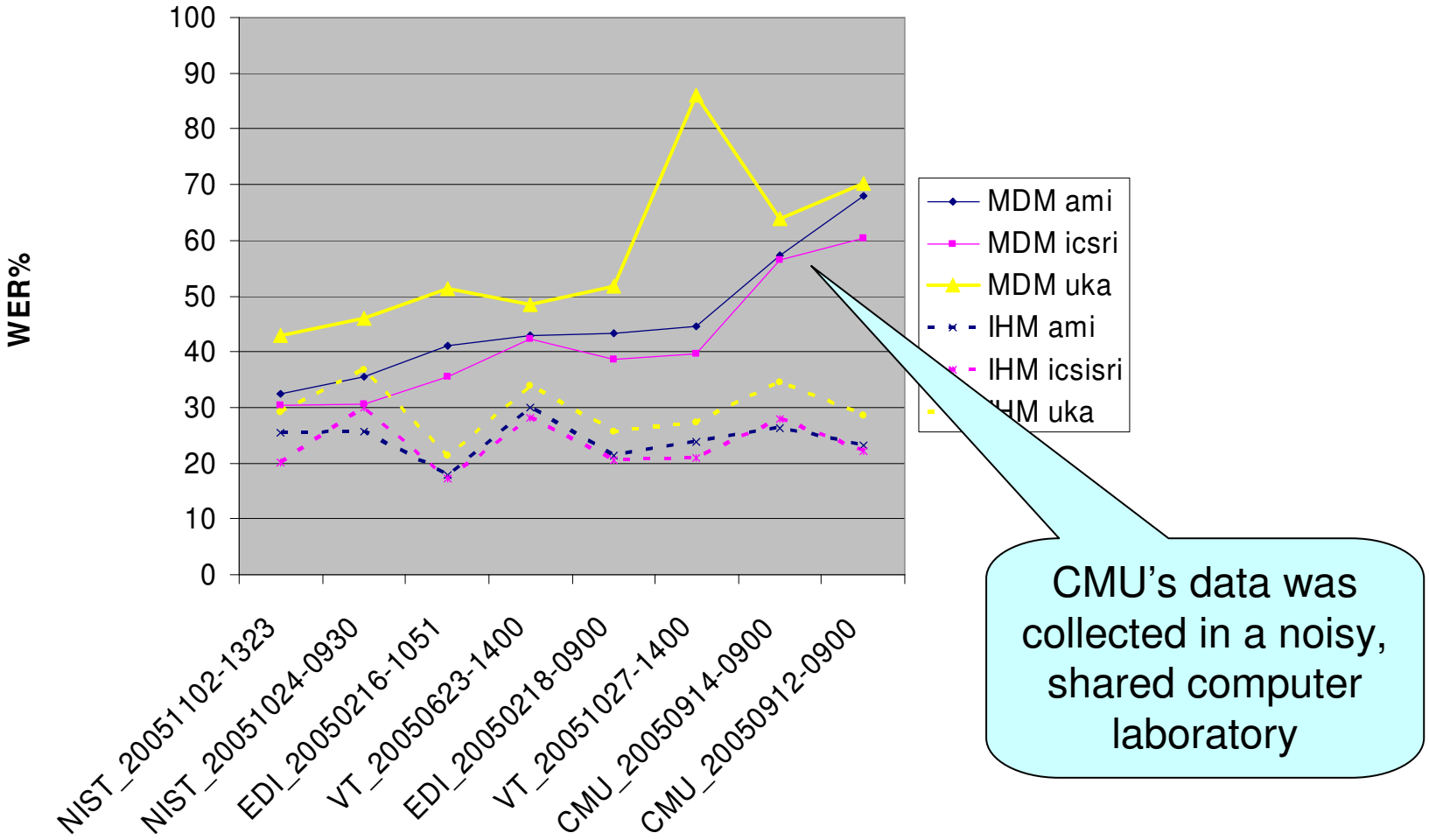


- Current Overlap<=4 WERs aren't compatible with '05 but they are definitely higher than '05
- RT-06S set looks more difficult in terms of acoustic challenge, but not language

NIST National Institute of Standards and Technology

Conference Data by Meeting ID

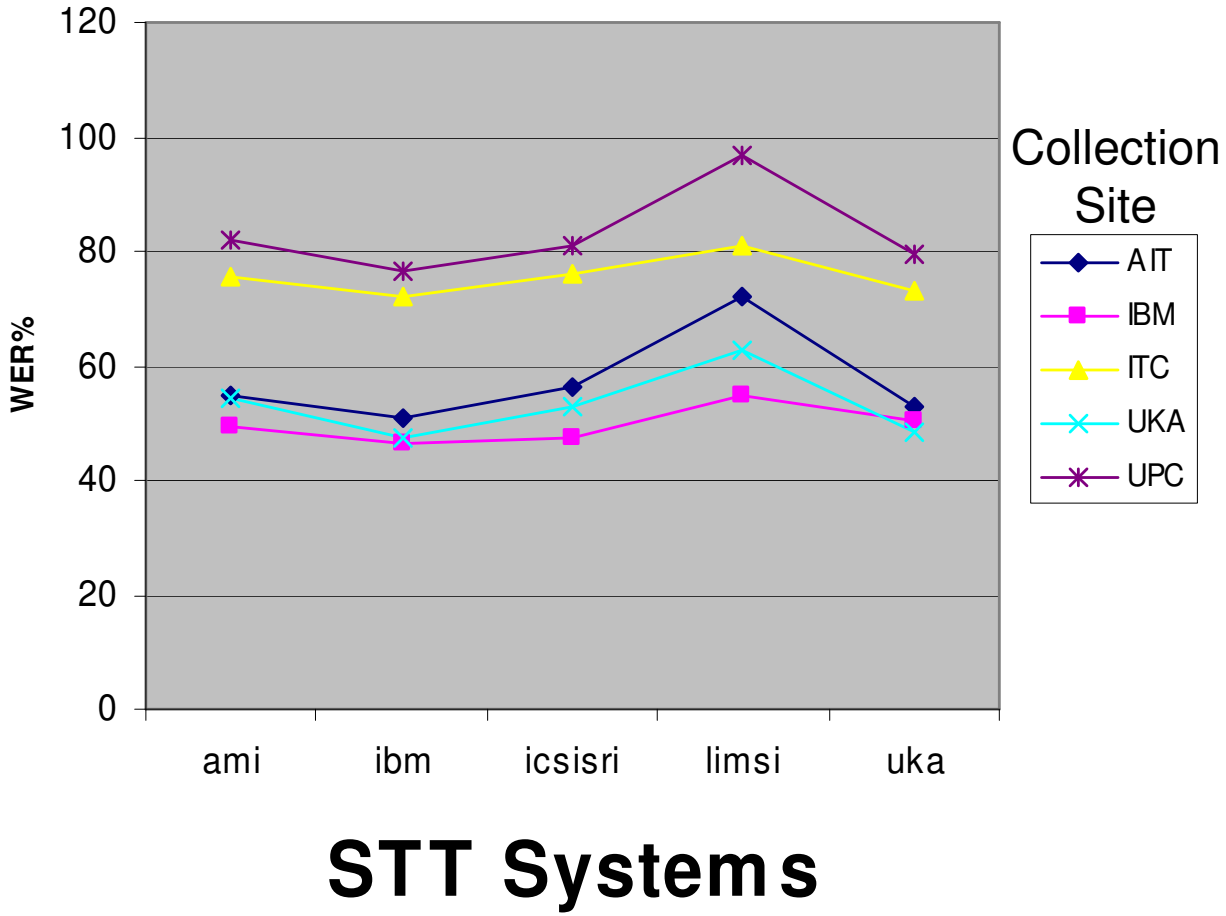
IHM and MDM Results for Primary Systems



NIST National Institute of Standards and Technology

Lecture Data by collection site

Current MDM Results for Primary Systems



Horizontal separation

- Background acoustics
- Accent/Training Data

NIST National Institute of Standards and Technology

Next Steps/Conclusions

- We need to plan in time for reference fixes in the future
- A decision needs to be made to either have every one run the missing TNO distant mic data for the conference data or leave it out of the official scores
- Scores need to be finalized
  - Lecture IHM results need re-scored
  - All distant mic conditions need to be re-scored
    - Conference error rates will increase
    - Lecture data will not change much
- Schedule for system description papers needs to be softened